Privacy Preserving Data Mining

Dheeraj Jayachandra
School of Computer Science
University of Windsor
jayachad@uwindsor.ca

Subodh Dwaraka Rao Pandraju
School of Computer Science
University of Windsor
pandrajs@uwindsor.ca

Vamshy Pagadala
School of Computer Science
University of Windsor
pagadal1@uwindsor.ca

Harshavardhan Naidu Gangavarapu
School of Computer Science
University of Windsor
gangavah@uwindsor.ca

***Abstract*—Nowadays, consumers are extremely using the digital technology, and the generated data is used by many companies to provide better service and engagement to the them. This not only benefits the consumers but also gives high profits to the companies. These companies and enterprises have a huge responsibility to keep the generated data safe from hackers and employees of organizations who can steal and have access to sensitive information such as bank details. Protecting privacy is highly important as it ensures dignity and safety to humans. In the project we used privacy preservation data mining methods to protect valuable data. We used K-means clustering algorithm, K-Anonymity and Cryptography in order to achieve the privacy on a dataset, which contains extremely sensitive attributes like Credit card details, income, IP address and age. We also compared these three Data Mining techniques to find out which one is the most suitable one.**

***Index Terms*—Privacy Preserving, K-means, anonymity, cryptography, data mining**

I. Introduction

In this digitized age, data mining has been advantageous to countless services such as banking, healthcare industry, and commerce. Nevertheless, the collected data is private information, which raises privacy issues. The information scope of privacy deals with collection and handling of sensitive data, and the main idea behind it is to have control on them. To safeguard from data leaks, privacy preservation methods using data mining are used to protect an individual's exposure, and it can be done by modifying the original data. Data mining is a fundamental process of automatically discovering big data to provide meaningful insights and relations, which is hard for humans to perform and analyze manually. There are various algorithms to mine the data, such as K-Means Clustering, Cryptography, K-Anonymity, etc. The mined data can be of any formats like clusters, patterns or classified data. The privacy protection data mining deals on how to assure that the information such as debit/credit card numbers, account login details, address, and salary details does not get disclosed while generating meaningful insights and relations in the data.

II. Related Work

With the rapid increase in data mining at different sectors such as healthcare, communication and business, user data is in the hands of more people every day. This data often includes sensitive data about individuals that isn't always required. Privacy preserving data mining is a technique where privacy can be implemented alongside data mining techniques.

PPDM algorithms can be classified into two types [1]:

- Data hiding: Algorithms that hide sensitive data such as name, age and SIN are removed or hidden before data mining techniques are applied. [2] [3] [4]
- Rule hiding: These algorithms hide or remove sensitive information that's derived after applying data mining techniques. [5] [6]

Most PPDM techniques are based on association rule mining, classification and clustering as these are the most common data mining techniques. In this paper, we'll focus on three methods introduced previously and compare them to analyze which works best for preserving privacy under different scenarios.

First is PPDM on k-means clustering algorithm [7]. This is a rule hiding technique where we hide data that's been generated from k-means clustering. The second method is k-anonymity which was introduced by Sweeney [8]. It requires each record in a dataset to be indistinguishable from at least k-1 other records in the dataset. It is a form of data hiding.

Thirdly, we have cryptography. Kantarcioĝlu and Clifton [9] were one of the first to suggest cryptography to protect user data back when data mining was mainly used by government agencies to deal with national security concerns. This is also a form of data hiding.

Based on this research, we aimed to preserve privacy using anonymization approach and draw a comparison between these three techniques and decide which technique is most effective for preserving privacy.

III. Proposed Model

To evaluate the amount of data that is preserved without any leaks in any dataset we chose to use three Data Mining techniques, K-Means Clustering, K-anonymity, Cryptography and draw comparisons between these to show the best and efficient technique. We used these techniques to achieve the privacy on an autogenerated dataset, which has 1000 rows and 8 columns. It contains sensitive attributes like “Age”, “Income”, “Credit Card Details”.

A. K-means Clustering

K-Means is an unsupervised learning algorithm, this will categorize the data based on the similarities, and find patterns in the dataset which are represented by k-clusters. The clusters are the datapoints grouped together based on their similarities [10].

Our very first step is to classify the attributes from the dataset as “Quasi Identifier”, “Sensitive” and “Removal”. For the removal, one attribute column which requires to be completely removed otherwise it would reveal the identity. For our dataset we considered the attribute that must be removed was “Name”.

In the next step, we identify the clusters using K-Means approach where we set the ‘k’ value which tells into how many clusters we want the dataset to divided according to the k value. After this we performed anonymization on the ‘Income’ column by finding the mean for the income column which gave us the privacy preserved dataset.

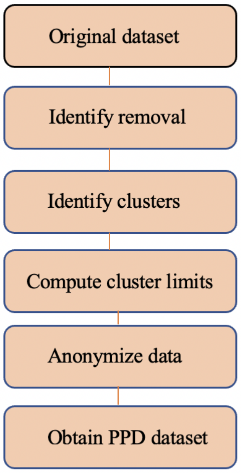

Original dataset
Identify removal
Identify clusters
Compute cluster limits
Anonymize data
Obtain PPD dataset

Fig. 1. Workflow of K-means

B. K-Anonymity

K – Anonymity a privacy model that is frequently employed to preserve the privacy of the data subjects in data sharing situations and the assurances that k-anonymity can offer when data is anonymized. To address the possibility of re-identification of anonymized data by linking to other datasets, the core concept of K-anonymity was developed. To achieve k-anonymity, the dataset must contain at least k individuals who share the set of qualities that may be used to identify everyone. One way to think about K-anonymity is as a ”hiding in the crowd” guarantee: if every person is a part of a bigger group, then any of the records in this group could be associated with a single person. K-Anonymity data anonymization technique that relies on suppression and generalization [11] [12].

Here we took the same autogenerated dataset which has 1000 rows and 8 columns to achieve K- anonymity. In the next step we identified the private information. With this information we targeted the sensitive columns and anonymized these columns using suppression and generalization. In the last step we described the columns with their respective anonymization types.

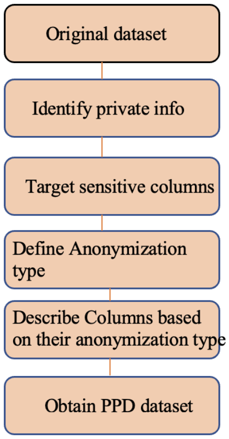

Original dataset
Identify private info
Target sensitive columns
Define Anonymization type
Describe Columns based on their anonymization type
Obtain PPD dataset

Fig. 2. Workflow of K-anonymity

C. Cryptography

Utilizing codes to secure communications and information so that only the intended audience can decipher and process it is known as cryptography. As a result, information access that isn’t authorised is prevented [13]. In cryptography, plaintext is converted into ciphertext and ciphertext is converted back into plaintext. We can encrypt and decrypt data using the cryptography module that Python supports. The cryptography package’s fernet module includes functions for the creation of the key, encrypting plaintext into ciphertext, and decrypting ciphertext back into plaintext using the encrypt and decrypt methods, respectively. Without the key, the fernet module ensures that data encrypted with it cannot be altered or read in any way [14].

To achieve cryptography, we used to same autogenerated dataset 1000 rows and 8 columns. Using the fernet module we can generate keys by using its methods i.e., generate_key() this method generates a fernet key.

In the next step, we convert the whole data in the dataset to type string to encrypt the data. We encrypt the converted data, and it can only be decrypted with the key.

IV. Results

A. K-means Clustering

In this section we will be focussing on the results of our data mining techniques and draw comparisons between them to showcase the efficient technique that can be used to achieve the privacy. Here we will show each technique with its anonymized dataset, first let us look at K-means clustering results i.e., Figure 4 and Figure 5

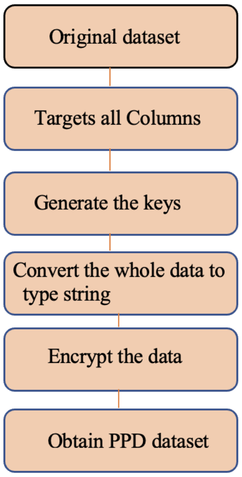
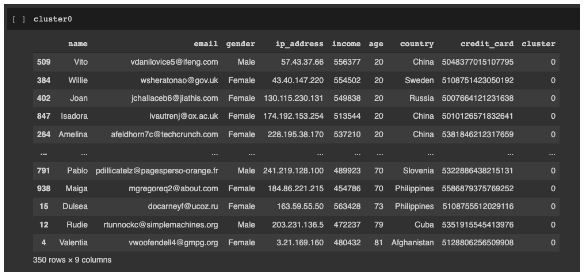


Fig. 3. Workflow of Cryptography

In Figure 4, the data is not anonymized, but it is formed into clusters and is ready for the anonymization, here we chose the attribute "Income" to be anonymized. In Figure 5, we performed the anonymization on the 'Income' column by finding the mean of each cluster which gives the privacy preserved dataset. As we can see here, we are able to anonymize only one attribute 'income' but the dataset is still prone to attack so the privacy of users information is still at stake.

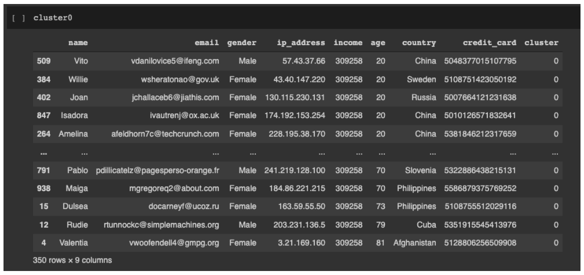


```
[ ] cluster0
```

	name	email	gender	ip_address	income	age	country	credit_card	cluster
509	Vito	vdanilovice5@ifeng.com	Male	57.43.37.66	556377	20	China	5048377015107795	0
384	Willie	wsheratonao@gov.uk	Female	43.40.147.220	554502	20	Sweden	5108751423050192	0
402	Joan	jchallaceb6@jiathis.com	Female	130.115.230.131	549838	20	Russia	5007664121231638	0
847	Isadora	ivautrenj@ox.ac.uk	Female	174.192.153.254	513544	20	China	5010126571832641	0
264	Amelina	afeldhorn7c@techcrunch.com	Female	228.195.38.170	537210	20	China	5381846212317659	0
...	...	...	...	...	...	...	...	...	...
791	Pablo	pdillicatelz@pagesperso-orange.fr	Male	241.219.128.100	489923	70	Slovenia	5322886438215131	0
938	Maiga	mgregoreq2@about.com	Female	184.86.221.215	454786	70	Philippines	5586879375769252	0
15	Dulsea	docarneyf@ucoz.ru	Female	163.59.55.50	563428	73	Philippines	5108755512029116	0
12	Rudie	rtunnockc@simplemachines.org	Male	203.231.136.5	472237	79	Cuba	5351915545413976	0
4	Valentia	vwoofendell4@gmpg.org	Female	3.21.169.160	480432	81	Afghanistan	5128806256509908	0

350 rows × 9 columns

Fig. 4. K-Means dataset before anonymizing



```
[ ] cluster0
```

	name	email	gender	ip_address	income	age	country	credit_card	cluster
509	Vito	vdanilovice5@ifeng.com	Male	57.43.37.66	309258	20	China	5048377015107795	0
384	Willie	wsheratonao@gov.uk	Female	43.40.147.220	309258	20	Sweden	5108751423050192	0
402	Joan	jchallaceb6@jiathis.com	Female	130.115.230.131	309258	20	Russia	5007664121231638	0
847	Isadora	ivautrenj@ox.ac.uk	Female	174.192.153.254	309258	20	China	5010126571832641	0
264	Amelina	afeldhorn7c@techcrunch.com	Female	228.195.38.170	309258	20	China	5381846212317659	0
...	...	...	...	...	...	...	...	...	...
791	Pablo	pdillicatelz@pagesperso-orange.fr	Male	241.219.128.100	309258	70	Slovenia	5322886438215131	0
938	Maiga	mgregoreq2@about.com	Female	184.86.221.215	309258	70	Philippines	5586879375769252	0
15	Dulsea	docarneyf@ucoz.ru	Female	163.59.55.50	309258	73	Philippines	5108755512029116	0
12	Rudie	rtunnockc@simplemachines.org	Male	203.231.136.5	309258	79	Cuba	5351915545413976	0
4	Valentia	vwoofendell4@gmpg.org	Female	3.21.169.160	309258	81	Afghanistan	5128806256509908	0

350 rows × 9 columns

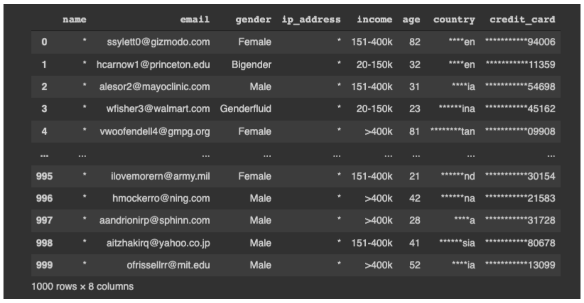
Fig. 5. K-Means dataset after anonymizing

B. K-Anonymity

Now let us see the K-anonymity results. In the original dataset, the data is not anonymized, and we identified the private information and targeted those specific columns that needs to anonymize. In K- anonymity the anonymization relies on suppression and generalization so the targeted columns will be suppressed and generalized accordingly.

In Figure 6, the dataset is anonymized based on their type i.e.., suppressed, semi suppressed and generalization. The columns whose data is completely hidden is the type of suppression, the columns whose data is partially hidden is semi suppressed and the column 'income' is generalized by giving an estimate value.

So, as we can see all the sensitive columns in this dataset are hidden and the chance of attacking this dataset is very less.



	name	email	gender	ip_address	income	age	country	credit_card
0	*	ssylett0@gizmodo.com	Female	*	151-400k	82	****en	***********94006
1	*	hcarnow1@princeton.edu	Bigender	*	20-150k	32	****en	***********11359
2	*	alesor2@mayoclinic.com	Male	*	151-400k	31	****ia	***********54698
3	*	wfisher3@walmart.com	Genderfluid	*	20-150k	23	******ina	***********45162
4	*	vwoofendell4@gmpg.org	Female	*	>400k	81	********tan	***********09908
...	...	...	...	...	...	...	...	...
995	*	ilovemorern@army.mil	Female	*	151-400k	21	******nd	***********30154
996	*	hmockerro@ning.com	Male	*	>400k	42	******na	***********21583
997	*	aandrionirp@sphinn.com	Male	*	>400k	28	****a	***********31728
998	*	aitzhakirq@yahoo.co.jp	Male	*	151-400k	41	******sia	***********80678
999	*	ofrissellrr@mit.edu	Male	*	>400k	52	****ia	***********13099

1000 rows × 8 columns

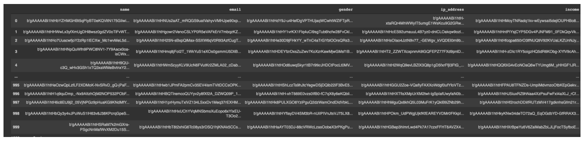
Fig. 6. K-anonymity dataset after anonymizing

C. Cryptography

Now let us look at the Cryptography technique results in Figure 7.

In Figure 7, we can see the dataset that is completely encrypted using the cryptography technique. Using the python module fernet we generated a key and changed the data in the dataset to type string to perform encryption. The data you see in the Figure 7 is after the encryption. This data can only be decrypted with the key.

However, the results of a calculation are not secure with cryptography. Instead, it stops computation-related privacy leakage. As a result, it falls short of offering a comprehensive solution to the issue of privacy-preserving data mining.



	name	email	gender	ip_address	income
0	b'gAAAAABi1htHbYZHMGHBt5qPlyBT0eKQVitN175Gliwi...	b'gAAAAABi1htHNUs2aAT_mRQGS9uatVahyvVMHJpe90xp...	b'gAAAAABi1htHdY9J-u4HefDgVPTHUjaqWCwhtWZIFTpR...	b'gAAAAABi1htH-xtaRQr4MhWWyIT5umgE1WsKcu9G2GRw...	b'gAAAAABi1htHMoyTNRadq1kv-wEywsaI5dejlOUPHBo8...
1	b'gAAAAABi1htHHWwLx3yfXmUgDHl8wsz0gdZV0oTlFvWT...	b'gAAAAABi1htHgowr2VenoCSLYP0R5eVAFkErV7mbqzKZ...	b'gAAAAABi1htHY1vrKX1Flq4uCl9sgTub6hcfw28FvCkl...	b'gAAAAABi1htHcES92umauuL4B7yz0-dniCLOakqw9ozl...	b'gAAAAABi1htHntPiYsjCYP5DGv4PJNFM91_0FDkQqvVk...
2	b'gAAAAABi1htHo7UuacwfjoY2zRp1lECXw_Mc1wvMeL5d...	b'gAAAAABi1htH25N-sBxpQiMax-SSXW6qV0R93AzfCugT...	b'gAAAAABi1htHe30D9jFHkYY_wTnC4aTrGYbIOOrsQRs3...	b'gAAAAABi1htHsOlao4JztN9v77_-GEWgv_kVQDEI0m9b...	b'gAAAAABi1htHfcqpa650rO3lfNUQ9V92KVUaLKZUnNJy...
3	b'gAAAAABi1htHNpQuWfn9PWC8NV1-7Y9Aacx0oa-isCWx...	b'gAAAAABi1htHmq8jFol2T_1WkYuS1eXOs0gsmmU5DlIB...	b'gAAAAABi1htHDEYfzrDxsZuZwv7KoXzrKawMjwGMsf1B...	b'gAAAAABi1htHT2_ZZWTXcxpnmA9lGQFEPZ77FXd9pnlD...	b'gAAAAABi1htH-zDlc1RY5oigzHQ5dR6KObg-XYfV9cAh...
4	b'gAAAAABi1htH9QU-c3Q_wHv3GSh1xTQ3xaWMeBvtnxY2...	b'gAAAAABi1htHWmScyyKLV9UcN6FVutKr2ZMLA02_cDab...	b'gAAAAABi1htHOd6uwqSkyr1tB7r9tkrJHDCtPocLt0MV...	b'gAAAAABi1htH2WqQ9evLBZlX3Q8p1gD5fxrFfjl3FtG_...	b'gAAAAABi1htHQQf0iG4vEoNOaQ6wTYUmg6M_yHHGF1JR...
...	...	...	...	...	...
995	b'gAAAAABi1htHeOzwQpLzfLF2XDMoK-NvSRvD_gQ-jPaF...	b'gAAAAABi1htHwb1JPmFA2pmCv35EV4lsmTVitDCCsOPK...	b'gAAAAABi1htHShLczTs9hJtcYegwDSjDQib22iF3BvE5...	b'gAAAAABi1htHGGU2Zoe-VQaflyFAXXoWdgf0ufYfoVTz...	b'gAAAAABi1htHFPAU8TFNZDs-UmplMdvmzoOtbKEpGekv...
996	b'gAAAAABi1htH1q9quDmp_-Nx9AltzhOjM2NOP532PObF...	b'gAAAAABi1htHBQTlremcoQUYyv2y8fXf2A_DZWQ09P_1...	b'gAAAAABi1htH-xhT669E0vzs0lfB0-fC7qX6y6eKQbli...	b'gAAAAABi1htHXTfisXMpUrXM2twt-IgSplafUwyfaN0b...	b'gAAAAABi1htHJSiuxRZ9hNaa4XzPsuFwFxl4aXLJ_rCf...
997	b'gAAAAABi1htH6c8EtJ9j2_05VjNPGz9pHuaKG9KNdMfY...	b'gAAAAABi1htH1prHymuTxlVZ134LSxxDv1Meq37rEXHM...	b'gAAAAABi1htHk8PULXGO8YziPguQ2dzWamOndDldVbkL...	b'gAAAAABi1htHhMguQx8khQ5L03MuFrK1yQklB9ZNb29h...	b'gAAAAABi1htHf2rochDDiifRUTzMV417gdknhsGfm21r...
998	b'gAAAAABi1htHbQy3y4vJPuWuS1lH63v8J38KPcrqGpeS...	b'gAAAAABi1htHxUChYVqMNSbmsXuEopo6xYIxEU-T3Oo2...	b'gAAAAABi1htHYflayDV4SM3bR-nUilPIVnJtsVJ75LX8...	b'gAAAAABi1htHPOkm_UdPWgjUjkfKfEAREYVDMr0FKkpl...	b'gAAAAABi1htHkyKNw34deTO72aQ_EqOGdbYD-GfRRAX3...
999	b'gAAAAABi1htHSRaM7k2mGXra-PSgcNnMafWvXM2Du15S...	b'gAAAAABi1htHbT8t2xhiG6Tc08ys3rDSQ1hjKNAk5CCs...	b'gAAAAABi1htHaAYT03DJ-88cVRWcLzasOobeX3rPKgPu...	b'gAAAAABi1htHGBep3hlmrLwd4Pk7A17czxFFHT6AVZX4...	b'gAAAAABi1htHXrBpeYu6V6ZalMabZbLJLjFocTSyfboE...

Fig. 7. Dataset after encryption

V. Limitations

In this research we used three different techniques to compare effectiveness. Every technique has its limitations and here we explained few potential limitations that might affect the effectiveness of the technique.

For K-means clustering it is to specify the k value i.e.., the number of clusters in the beginning it is difficult to choose the appropriate k value for some datasets, and it can handle only the numerical data which is a concern where in a dataset there would be different attributes that are sensitive.

K-anonymity assumes that each private attribute takes values that are consistent across its domain, or that the frequencies of attribute's various values are comparable also the anonymization approaches are prone to homogeneity attack and background knowledge.

And in the case of Cryptography there might be key leakage, phishing attacks, and other major drawbacks are even a valid user may find it challenging to access secure encrypted, authenticated, and digitally signed information during a moment when decisions are most important. One of challenge we faced during this project is to achieve anonymization approach on other techniques. It was difficult for us to understand few libraries and their functions that are crucial for our project.

VI. Conclusion and Future Work

To conclude, based on our research and analysis we compared three data mining techniques K-Means Clustering, K-Anonymity, Cryptography. We applied anonymization approach on all the three techniques to find the efficient technique that is giving the best privacy preserved dataset and on analysing the results, K- Anonymity gave the best results out of the other two techniques. K- Anonymity anonymized most of the sensitive data using suppression and generalization. Moreover, we intend to formally test the entire spectrum of PPDM algorithms as part of our ongoing research.

References

[1] Wu, Xiaodan Chu, Chao Wang, Yunfeng Liu, Fengli Yue, Dianmin. (2007). Privacy Preserving Data Mining Research: Current Status and Key Issues. 4489. 762-772. 10.1007/978-3-540-72588-6125.

[2] Islam, Md Brankovic, Ljiljana. (2004). A Framework for Privacy Preserving Classification in Data Mining.

[3] S. R. Oliveira and O. R. Zaïane, "Achieving privacy preservation when sharing data for clustering," Lecture Notes in Computer Science, pp. 67–82, 2004.

[4] Shariq J. Rizvi and Jayant R. Haritsa. 2002. Maintaining data privacy in association rule mining. In Proceedings of the 28th international conference on Very Large Data Bases (VLDB '02). VLDB Endowment, 682–693.

[5] Oliveira, Stanley Zaïane, Osmar. (2003). Protecting Sensitive Knowledge By Data Sanitization.. 613-616. 10.1109/ICDM.2003.1250990.

[6] Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yucel Saygin, and Elena Dasseni. 2004. Association Rule Hiding. IEEE Trans. on Knowl. and Data Eng. 16, 4 (April 2004), 434–447. https://doi.org/10.1109/TKDE.2004.1269668

[7] C.-Y. Lin, "A reversible privacy-preserving clustering technique based on K-means algorithm," Applied Soft Computing, vol. 87, p. 105995, 2020.

[8] Sweeney, L.. (2002). k-Anonymity: A Model for Protecting Privacy. IEEE Security and Privacy. 10. 1-14.

[9] M. Kantarcioĝlu and C. Clifton, "Assuring privacy when big brother is watching," Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery - DMKD '03, 2003.

[10] https://www.inovex.de/de/blog/disadvantages-of-k-means-clustering/

[11] https://www.privitar.com/blog/k-anonymity-an-introduction/

[12] Domingo-Ferrer, Josep Torra, Vicenç. (2008). A Critique of k-Anonymity and Some of Its Enhancements. ARES 2008 - 3rd International Conference on Availability, Security, and Reliability, Proceedings. 990-993. 10.1109/ARES.2008.97.

[13] https://www.geeksforgeeks.org/cryptography-and-its-types/

[14] https://www.geeksforgeeks.org/fernet-symmetric-encryption-using-cryptography-module-in-python/